Analítica Avanzada – Energía

Primer Informe

Carmen Cecilia Sánchez Zuleta(UDM)

Roxana Alemán(UAJMS)

Juan Pablo Fernández Gutiérrez(UDM)

Fabio Humberto Sepúlveda Murillo(UDM)

Pérdidas no técnicas
Estado de las Bases de Datos de la Empresa SETAR-Bolivia a diciembre de 2019

Universidad de Medellín

Facultad de Ciencias Básicas

2020

Tabla de contenido

Introducción 3

1. Trazabilidad de los Datos 3

1.1. Base de datos Clientes 4

1.2. Base de Datos de los Usuarios Infractores 7

2. Análisis de la Data de Consumo 10

2.1 Consumo JUL a DIC 2016 10

2.2 Consumo ENE a JUN 2017 11

2.3 Consumo JUL a DIC 2017 13

2.4 Consumo ENE a JUN 2018 15

2.5 Consumo JUL a DIC 2018 17

2.6 Consumo ENE a JUN 2019 18

3. Estado de las Bases de Datos Espaciales 21

4. Piloto distrito 3 - Operativos de revisión de medidores 22

Conclusiones y Recomendaciones 29

Introducción

Enmarcados en una propuesta que tiene como propósito utilizar las herramientas de la analítica para clasificar y caracterizar los usuarios de la empresa SETAR, de tal manera que, a partir de los históricos de la entidad, se puedan detectar diferentes focos de las pérdidas no técnicas que sufre la empresa, se presenta en este documento los resultados de la primera etapa que todo proyecto de esta índole debe incluir: la limpieza de los datos, y con él, el estado de la data.

Se sigue entonces que, en busca de lograr los objetivos del proyecto en esta primera etapa de su desarrollo, se ha adelantado un análisis exploratorio de los datos, haciendo uso tanto de la estadística descriptiva convencional, como de la estadística espacial, y otras técnicas para analizar el estado de las bases de datos, observar algún comportamiento de las variables involucradas, y determinar, a partir de la data, la pertinencia de metodología seleccionada para el trabajo.

A continuación, en cuatro secciones, se presentan los hallazgos encontrados en el desarrollo de la primera etapa del proyecto.

1. Trazabilidad de los Datos

La trazabilidad de los datos está relacionada con la posibilidad de cruzar la información disponible en diferentes bases de datos. Disponer de una buena trazabilidad es fundamental cuando los datos proceden de diferentes fuentes, dado que con frecuencia se requiere fusionar información de uno de los archivos con los de otro.

Para esta primera tarea se contó con las siguientes bases de datos:

- **Clientes:** Es una data que hace una descripción a partir de seis variables de 78158 usuarios de la empresa SETAR.

- **Usuarios Infractores:** Es una data que relaciona los usuarios infractores identificados con las multas impuestas y el monto recuperado. Presenta además el detalle de la sanción y el estado del proceso.

- **Consumos:** Base datos constituida por los informes mensuales del consumo dado en KW. La dimensión cambia de semestre a semestre.

- **Medidores Georreferenciados:** Esta es la base de datos que contiene la ubicación georrefenciada de los medidores que la empresa tiene registrada en toda la ciudad.

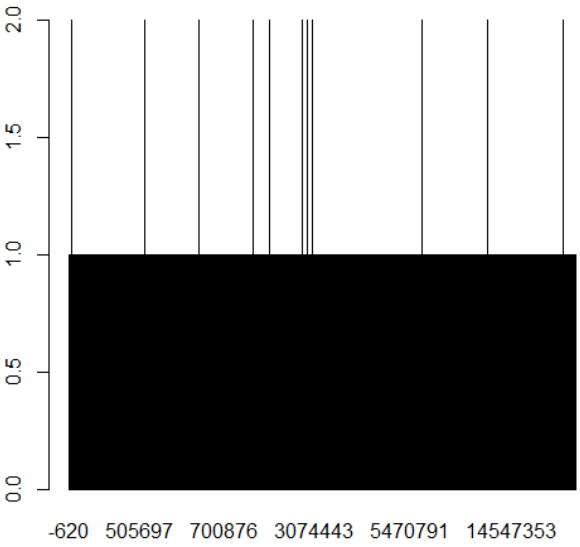
El uso de estas tres bases de datos con fuentes diferentes demanda la necesidad de que la información registrada sea lo suficientemente clara y unificada en cada una de ellas, de tal manera que permita cruzar la información que sea necesaria entre las bases.

1.1. Base de datos Clientes

A continuación, se describe el estado de la data entregada por la empresa SETAR (Tarija, Bolivia) a fecha de julio de 2019, específicamente este archivo se centra en la información encontrada en la exploración de los datos en relación con los usuarios, su identificación y el medidor asociado, así como la trazabilidad de las datas.

Respecto a esta base de datos los principales hallazgos son:

1. Con relación al medidor se ha encontrado que existen usuarios diferentes cuyos medidores están referenciados con el mismo número y con direcciones diferentes (figura 1).




```
> FrecMedi
MEDIDOR
      3104     551346     679524     848642    1006370    2703829
         2          2          2          2          2          2
   2890671    3072778    3073488    5522649   13827926  222267777
         2          2          2          2          2          2
```

```
> MededREP
       MEDIDOR CLIENTE                          NOMBRE                                    DIRECCION            AREA
380       3104    7391                  ZELAYA  ERNESTINA                                 COCHABAMBA 0          TARIJA DOM
381       3104   85600       ANAGUA CRUZ DE COLLARANI VICENTA                          B/TABLADITA II 0      TARIJA DOM
11524   551346    6639              LLANOS  MIGUEL-COMERCIAL          AV. DOMINGO PAZ/ DANIEL CAMPOS Y COLON 173 TARIJA G-1
11525   551346    8561                 VALDEZ MENDEZ GERMAN                           HUGO LOPEZ DOLZ 1256  TARIJA DOM
19903   679524   12772            MARQUEZ  VERA LUCIA  VACA DE                    B/EL ROSEDAL C/Camatindi 3155 TARIJA DOM
19904   679524   60332                CASTRO CARVAJAL HERNAN                        CARAPARI CHUQUISACA 0 SAN LORENZO DOM
28197   848642   90963          LOPEZ CORTEZ CARLOS ALBERTO                  TURUMAYO, B/MIRADOR LOS PINOS 0      TARIJA DOM
28198   848642   90970        VELASQUEZ FIGUEROA JUAN GONZALO                        B/ANICETO ARCE 0       TARIJA DOM
30628  1006370   74027             BENITEZ IBARBOL SANTIAGO                                CALAMA 0 SAN LORENZO DOM
30629  1006370   74029                CHOQUE ARGOTA CARLOS                                 CALAMA 0 SAN LORENZO DOM
35794  2703829   42262               RIOS ALVAREZ JUAN JOSE                             B/BOLIVAR 0       TARIJA G-1
35795  2703829   90967       BALDIVIEZO CARDOZO JANETH MARTHA          B/TARIJA NUEVA- ZONA LAS BARRANCAS 0      TARIJA G-1
36510  2890671   46687              REARTE GARECA CANDELARIO 12 DE OCTUBRE Av. Circunvalacion (Frente posta Vil 0 TARIJA INP
36511  2890671   71920              REARTE GARECA CANDELARIO          ASENTAMIENTO TARIJEÑOS EN PROGRESO 0      TARIJA INP
37244  3072778   81699 JAIME DONAIRE VILLA                    B° BOLIVAR/ AV. HEROES DEL CHACO ESQUI. MERC. SUR 0 TARIJA G-1
37245  3072778   82933                MARTINEZ  MARIA ALBA                          YESERA NORTE 0        TARIJA DOM
37362  3073488   82758           SORUCO BALDIVIEZO JESUSA                          YESERA CENTRO 0       TARIJA DOM
37363  3073488   82790 CLUB DEPORTIVO YESERA CENTRO                                  YESRA CENTRO 0        TARIJA DOM
54134  5522649   52108                MEDRANO GUZMAN DOLLY                           LA TERMINAL 0         TARIJA G-1
54135  5522649   89083                VLACHOS  CHARALAMPOS                          TORRECILLAS 0         TARIJA G-1
64088 13827926   39791          LEON ORDOÑEZ PAULINA DINA                          B/LA FLORIDA 0        TARIJA DOM
64089 13827926   69896          LEON ORDOÑEZ PAULINA DINA          B° FLORIDA/C/ COLON Y 15 DE AGOSTO 0      TARIJA DOM
75825 222267777   2549                 BORDA V. HUGO RAMON              O'CONNOR/ BOLIVAR Y INGAVI 735        TARIJA DOM
75826 222267777  66556                ARMELLA  ANGELITO FAUD                             CRUCE 0 ACHI - EL PUENTE DOM
```

Figura 1. Anomalía en Medidores – un medidor para dos usuarios

De la figura 1 se puede observar como el número de un medidor ha sido asignado a dos usuarios diferentes en direcciones diferentes. (¿Qué significa esto en el caso real?, o ¿hay un error?).

2. Esta data proporcionada cuenta con un total de 78157 usuarios de los cuales 431 no tienen número de medidor asignado (¿Cómo se controla la energía suministrada a estos predios?).

3. De acuerdo con la información suministrada el 89% de los usuarios están registrados en la categoría “Domiciliaria” y un 8% aproximadamente en la categoría “G-1” (Comerciales), lo que implica que el 97% de los usuarios de la empresa se encuentran registrados en estas dos categorías (figura 2). (Es importante identificar en qué categoría se encuentran los grandes consumidores).

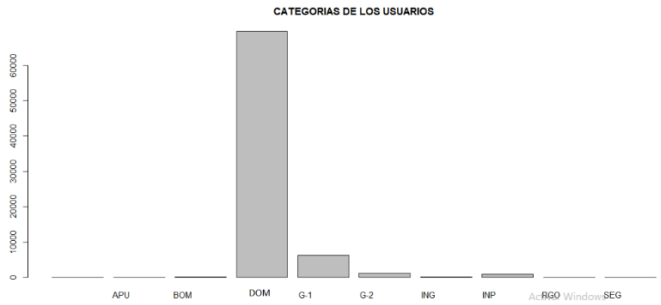


Figura 2. Categorías de los usuarios

4. Según lo que se ha acordado con el equipo de trabajo de Bolivia, la evaluación de las pérdidas se realizará sobre la ciudad de Tarija (inicialmente se había hablado de esto, posteriormente se estableció que se realizaría únicamente sobre el distrito 3.) pero de acuerdo a lo que muestra la distribución de los usuarios por área se están considerando otros sectores del estado (figura 3).

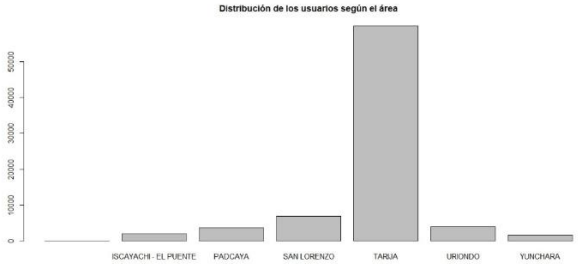
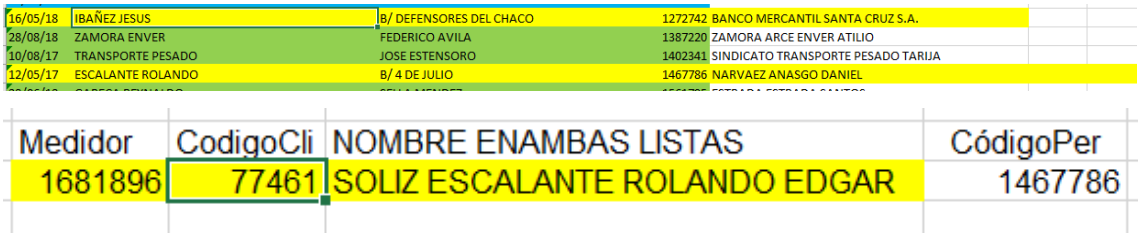


Figura 3. Distribución de los usuarios según área

Por otro lado, respecto a la consistencia de esta base de datos con las otras se encontró que:

1. La base de dato de clientes describe todos los clientes, sin embargo cuando tratamos de cruzar los usuarios de la data de pérdidas se encuentra que existe una inconsistencia entre los códigos de clientes, como se observa a continuación.



16/05/18	IBAÑEZ JESUS	B/ DEFENSORES DEL CHACO	1272742	BANCO MERCANTIL SANTA CRUZ S.A.
28/08/18	ZAMORA ENVER	FEDERICO AVILA	1387220	ZAMORA ARCE ENVER ATILIO
10/08/17	TRANSPORTE PESADO	JOSE ESTENSORO	1402341	SINDICATO TRANSPORTE PESADO TARIJA
12/05/17	ESCALANTE ROLANDO	B/ 4 DE JULIO	1467786	NARVAEZ ANASGO DANIEL

Medidor	CodigoCli	NOMBRE ENAMBAS LISTAS	CódigoPer
1681896	77461	SOLIZ ESCALANTE ROLANDO EDGAR	1467786

Figura 4. Inconsistencia de entre los códigos de los clientes entre las bases de datos

2. Con relación al ítem anterior percatamos que muchos de los códigos de usuarios en la data de “pérdidas” se relacionan con el “medidor”.

1.2. Base de Datos de los Usuarios Infractores

Con relación a la base de datos relacionada con los usuarios infractores, se encontraron diferentes inconsistencias que se describen a continuación.

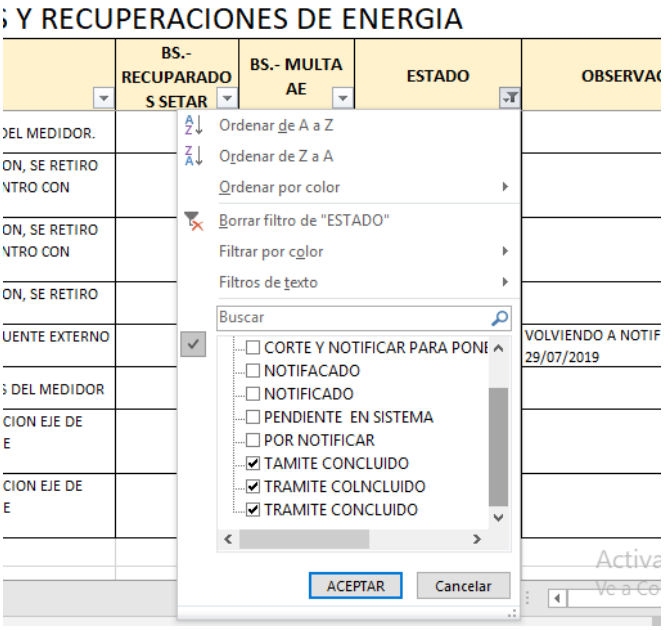
1. Esta base de datos cuenta con una variable no nombrada que presenta muchos datos faltantes, pero no se indica que representa las pocas medidas que aparecen en ella.

N°	N° CAS	FECHA INSPECCION	N° CLIENT	NOMBRE USUARIO	DETALLE	BS.- RECUPARADOS SETAR	BS.- MULTA AE	ESTADO	OBSERVACION	
58	2903	12/12/2018	13153	TOMAS EUGENIO MAMANI	SE ENCONTRO UNA DERIVACION ANTES DEL MEDIDOR.	1.761,22	196,40	TRAMITE CONCLUIDO		
68	2915	30/04/2019	2093	M. ROSA TRIGO DE CORTEZ	ALTERACION INSTRUMENTOS DE MEDICION, SE RETIRO PRECINTOS PLOMOS DUDOSO Y SE ENCONTRO CON ERROR DE -99,9%	1.457,26	197,80	TRAMITE CONCLUIDO		2
70	2917	8/05/2019	29859	WALTER JARAMILLO BARRIENTOS	ALTERACION INSTRUMENTOS DE MEDICION, SE RETIRO PRECINTOS PLOMOS DUDOSO Y SE ENCONTRO CON ERROR DE -99,9%	1.339,96	199,20	TRAMITE CONCLUIDO		3
71	2918	9/05/2019	35647	FAUSTO ARANCIBIA BARRIENTOS	ALTERACION INSTRUMENTOS DE MEDICION, SE RETIRO PRECINTOS PLOMOS ROTOS	1023,4	199,2	TRAMITE CONCLUIDO		4
75	2922	17/05/2019	21462	ARMINDA CARDOZO PACHECO	MEDIDOR CON ERROR DE -1,640% CON PUENTE EXTERNO CON ERROR DE -75,691 %	515,37	199,2	TRAMITE CONCLUIDO	VOLVIENDO A NOTIFICAR 29/07/2019	5
76	2923	19/06/2019	3881	CATALINA CATA DE ESPINDOLA	SE ENCONTRO CONEXIÓN DIRECTA ANTES DEL MEDIDOR	1119,63	199,2	TRAMITE CONCLUIDO		
77	2925	10/07/2019	3012	MARIANELA FERNANDEZ	ALTERACION DE INSTRUMENTO DE MEDICION EJE DE DISCO TORCIDO INTENCIONAL Y PLACA DE CARACTERISTICAS SUELTAS	1054,47	199,2	TRAMITE CONCLUIDO		6
77	2926	10/07/2019	3012	MARIANELA FERNANDEZ	ALTERACION DE INSTRUMENTO DE MEDICION EJE DE DISCO TORCIDO INTENCIONAL Y PLACA DE CARACTERISTICAS SUELTAS	863,04	199,2	TRAMITE CONCLUIDO		7

Figura 5. Tabla con datos faltantes

2. Además, en la figura anterior también se puede observar que en la columna de estado, el estado “Tramite Consumo” aparece escrito de tres formas diferentes, sucede los mismo con el estado “NOTIFICADO”.

3. Entendiéndose la columna “BS.- RECUPARADOS SETAR” como el dinero que ha sido recuperado por la empresa, y “BS.- MULTA AE” como la multa impuesta, se tendría la nueva variable “Resta” que establece la multa impuesta menos lo recuperado tendría que ser todo el tiempo positivo, pero no está pasando esto, qué información real están presentando estas dos variables?



```
> summary(infracciones1$Resta)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
-16990.00  -1733.00   -920.40  -1398.00     -6.99   3015.00         2
> |
```

Figura 6. Estadísticos descriptivos para la multa impuesta

4. La falta de concordancia en los ID de los usuarios es fundamental para el cruce de información. En el primer semestre del 2018 no existe el ID 13606-1, note además que aparecen dos usuarios con ese código (figura 7).

CARPETAS CON INFRACCIONES Y RECUPERACIONES DE ENERGIA

N°	N° CASO	FECHA INSPECCION	N° CLIENTE	NOMBRE USUARIO	DETALLE	BS.- RECUPARADOS SETAR	BS.- MULTA AE	ESTADO
19	2858	5/04/2018	13606-1	RAMIRO VARGAS - TERESA YUJRA	ALTERACION INSTUMENTO DE MEDICION SE ECONTRO CON ERROR DEL 66%	879,00	205,00	TRAMITE CONCLUIDO

	A	B	C	D	E
	MEDIDOR	CLIENT	NOMBRE	DIRECCION	AREA
4	828628	13606	VARGAS RAMIRO	B/SAN GERONIMO 0	TARIJA

Figura 7. Inconsistencia de los ID

5. Respecto a un resumen de la “¿La infracción” neta de los usuarios se observa que se tiene cuatro usuarios sin cliente registrado, y además surge la pregunta: ¿cómo se puede interpretar los ID de los usuarios dados con el menos uno? (figura 8).

```
> infracciones1$N..CLIENTE
 [1] 21088          71371          23037          10332
 [5] 42668          47724          8859           34728
 [9] 3289           5085           28983          9793
[13] 55223          5190           52551          C/ANGEL CALAVI
[17] 6647           13606-1        58091          7098-1
[21] 9408-1         71079          13154          9723
[25] 23444-1        1924           26773          703086-1
[29] 70386          25801          36945          23371
[33] 24295          1713           23696          640
[37] 14140          50895          33794          82161
[41] 34850          31904          8935           31905
[45] 36178          41246          84640          54260
[49] 10640          44943          32644          62613
[53] 6482           35155          13153          79882
[57] 18777          44382          23941          54422
[61] 10640          29132
[65] 2093           54582          29859          35647
[69] 58863          43376          29923          21462
[73] 3881           32539          3012           3012
[77] 79859          36024          43835
[81]
```

Figura 8. Registro de los clientes infractores

2. Análisis de la Data de Consumo

El siguiente análisis de la calidad de datos del consumo de divide en: Consumo JUL a DIC 2016, Consumo ENE a JUN 2017, Consumo JUL a DIC 2017, Consumo ENE a JUN 2018, Consumo JUL a DIC 2018, y Consumo ENE a JUN 2019 entregados por los colegas de la Universidad Juan Misael Saracho de Bolivia.

Los principales hallazgos en cada una de éstas son:

2.1 Consumo JUL a DIC 2016

1. Esta base contiene 8 columnas, donde la primera es el ID del cliente que está en un formato no estándar a recomendaciones en bases de datos, puesto que el ID es dado por un número que aumenta su cantidad de dígitos identificado con la etiqueta con “CLIENTE”. En la columna 8 identificada con “Estado” no tiene niveles faltantes.

	JUL_16	AGO_16	SEP_16	OCT_16	NOV_16	DIC_16
count	89657.000000	89657.000000	89657.000000	89657.000000	89657.000000	89657.000000
mean	140.205562	141.765984	141.165548	143.650981	139.569267	145.990188
std	2719.406529	2723.811177	2748.025799	2769.067166	2605.167945	2649.814244
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	50.000000	51.000000	51.000000	53.000000	55.000000	50.000000
75%	126.000000	128.000000	124.000000	128.000000	126.000000	134.000000
max	632542.000000	636663.000000	607245.000000	612710.000000	601005.000000	601176.000000

Figura 9. Base de datos para Consumo JUL a DIC 2016

2. Además, observamos en la figura 9 la siguiente situación con la presencia de valores 0 en todas estas variables de los consumos mensuales con al menos el 25% de los datos en todos los estados y todos los usuarios en cada mes de consumo. También se nota un cambio alto entre el valor del consumo del cuartil 75% y su máximo en cada uno de los meses.

3. Los estados se distribuyen de la siguiente manera ilustrado en un gráfico de torta en la figura 10.

De la figura 10, se visualiza que los niveles 'SUSPENDIDO', 'CORTADO', 'CORTE EN CAMPO' de la variable categórica 'Estado' es sólo un 11% de la cantidad de datos de la base de datos, esto es, la base de datos es desbalanceada.

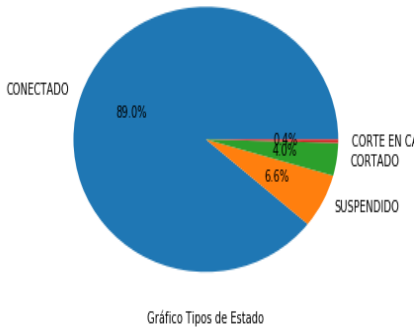


Figura 10. Estados de los usuarios 2016 semestre 2

4. Respecto a la identificación de los usuarios, la figura 11 nos muestra que la frecuencia de cada usuario es adecuada y nos muestra que existen ID sin asignación, los ID de los usuarios de la empresa no son continuos.

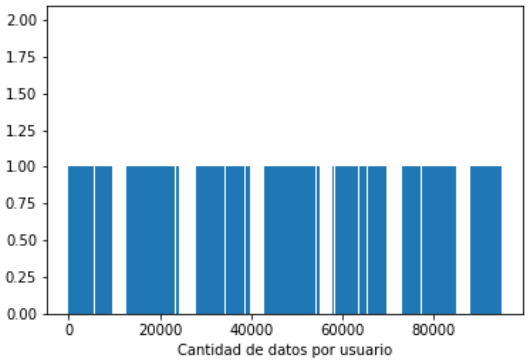


Figura 11. Distribución de la identificación de los usuarios

2.2 Consumo ENE a JUN 2017

1. Esta base de datos contiene 8 columnas, donde la primera es el ID del cliente que está en un formato no estándar a recomendaciones en bases de datos, puesto que el ID es dado por un número que aumenta su cantidad de dígitos identificado con la etiqueta con “CLIENTE”. En la columna 8 identificada con “Estado” no tiene niveles faltantes (figura 12).

	ENE_17	FEB_17	MAR_17	ABR_17	MAY_17	JUN_17
count	89658.000000	89658.000000	89658.000000	89658.000000	89658.000000	89658.000000
mean	163.971690	123.285240	142.625238	143.040031	145.323995	134.080334
std	2692.866723	2445.434896	2645.018021	2671.052165	2799.709947	2673.259592
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	68.000000	46.000000	57.000000	58.000000	59.000000	52.000000
75%	165.000000	106.000000	130.000000	132.000000	133.000000	116.000000
max	637892.000000	560570.000000	611530.000000	632962.000000	687498.000000	616506.000000

Figura 12. Base de datos para Consumo ENE a JUN 2017

2. Observamos de la figura 12 que la presencia de valores 0 en todas estas variables de los consumos mensuales son al menos el 25% de los datos en todos los estados y todos los usuarios en cada mes de consumo. También se nota un cambio alto entre el valor del consumo del cuartil 75% y su máximo en cada uno de los meses.

3. Los niveles 'SUSPENDIDO', 'CORTADO', 'CORTE EN CAMPO' de la variable categórica 'Estado' es sólo un 11% de la cantidad de datos de la base de datos, esto es, la base de datos es desbalanceada (figura 13).

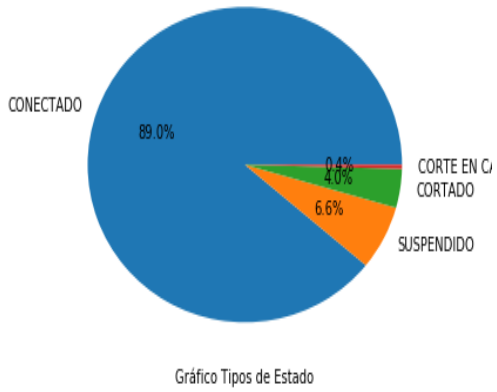


Figura 13. Estado de los usuarios 2017 semestre 1

4. La figura 14 muestra que la frecuencia de cada usuario es adecuada y nos muestra que existen ID sin asignación, los ID de los usuarios de la empresa no son continuos, sin embargo, se nota que la distribución de los usuarios cambia,

en especial en la región de los ID entre 40000 y 60000, indicando un movimiento de entra y salida de usuarios en el sistema, aunque la variación sea del aumento de un solo usuario más respecto a la cantidad del semestre anterior.

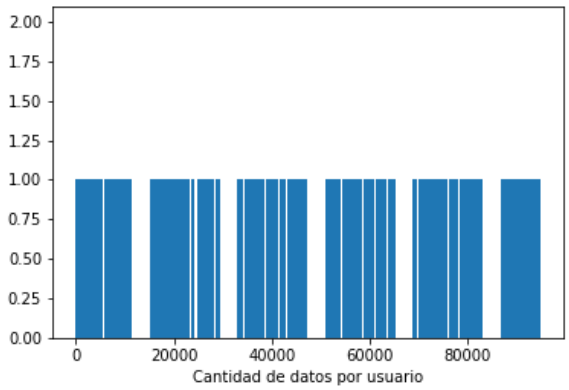


Figura 14. Distribución de la identificación de los usuarios

2.3 Consumo JUL a DIC 2017

1. La base de datos Consumo JUL a DIC 2017 en Excel contiene 8 columnas, donde la primera es el ID del cliente que está en un formato no estándar a recomendaciones en bases de datos, puesto que el ID es dado por un número que aumenta su cantidad de dígitos identificado con la etiqueta con “CLIENTE”. En la columna 8 identificada con “Estado” no tiene niveles faltantes.

	JUL_17	AG0_17	SEP_17	0CT_17	NOV_17	DIC_17
count	89611.000000	89611.000000	89611.000000	89611.000000	89611.000000	89611.000000
mean	141.954210	143.591025	148.461303	151.343467	157.089833	152.293730
std	2704.822908	2755.037776	2761.085814	3009.239443	2758.002464	2597.391338
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	1.000000	2.000000
50%	55.000000	57.000000	59.000000	60.000000	65.000000	62.000000
75%	126.000000	125.000000	132.000000	133.000000	139.000000	137.000000
max	634975.000000	634111.000000	633469.000000	714300.000000	625257.000000	587779.000000

Figura 15. Base de datos para Consumo JUL a DIC 2017

2. Dentro de las variables JUL_17, AGO_17, SEP_17, OCT_17, NOV_17 y DIC_17 de la base de datos Consumo JUL a DIC 2017, tampoco hay atributos perdidos, ni campos vacíos (figura 15).

3. Sin embargo, la siguiente situación con la presencia de valores 0 en todas estas variables de los consumos mensuales con al menos el 25% de los datos en todos los estados y todos los usuarios en cada mes de consumo. También se nota un cambio alto entre el valor del consumo del cuartil 75% y su máximo en cada uno de los meses.

4. Los niveles 'SUSPENDIDO', 'CORTADO', 'CORTE EN CAMPO' de la variable categórica 'Estado' es sólo un 11.7% de la cantidad de datos de la base de datos, esto es, la base de datos es desbalanceada (figura 16).

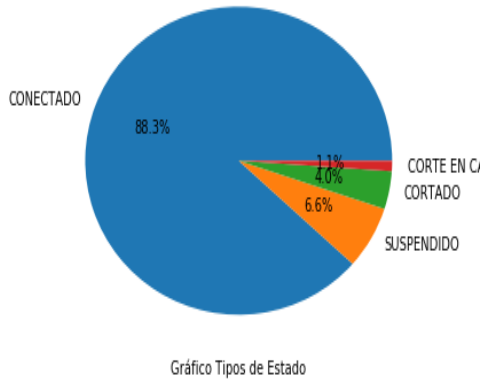


Figura 16. Estado de los usuarios 2017 semestre 2

5. La figura 17 nos muestra que la frecuencia de cada usuario es adecuada y nos muestra que existen ID sin asignación, los ID de los usuarios de la empresa no son continuos, sin embargo, se nota que la distribución de los usuarios cambia abruptamente con respecto a los dos semestres anteriores. Esto está indicando un movimiento de entra y salida de usuarios en el sistema, también podría generar cambios en los comportamientos de los usuarios.

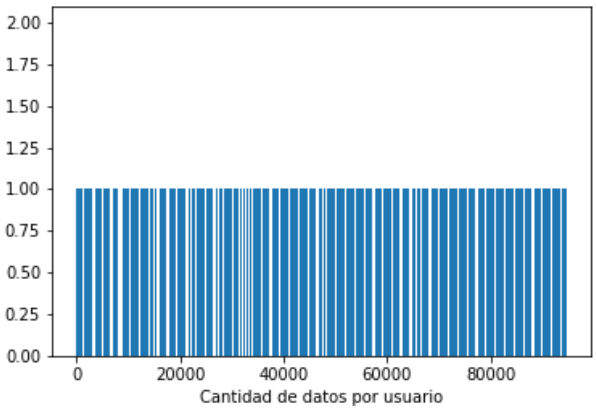


Figura 17. Distribución de la identificación de los usuarios

2.4 Consumo ENE a JUN 2018

1. La base de datos Consumo ENE a JUN 2018 contiene 8 columnas, donde la primera es el ID del cliente que está en un formato no estándar a recomendaciones en bases de datos, puesto que el ID es dado por un número que aumenta su cantidad de dígitos identificado con la etiqueta con “CLIENTE”. En la columna 8 identificada con “Estado” no tiene niveles faltantes.

	ENE_18	FEB_18	MAR_18	ABR_18	MAY_18	JUN_18
count	89611.000000	89611.000000	89611.000000	89611.000000	89611.000000	89611.000000
mean	157.634988	153.713056	145.166150	150.954986	154.698631	148.524465
std	2923.860703	3408.279139	2353.512318	2789.584644	2753.340228	2514.171982
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.000000	3.000000	2.000000	4.000000	5.000000	4.000000
50%	63.000000	65.000000	62.000000	62.000000	66.000000	60.000000
75%	141.000000	138.000000	131.000000	134.000000	140.000000	131.000000
max	635628.000000	872909.000000	492347.000000	658446.000000	652616.000000	555156.000000

Figura 18. Base de datos para Consumo ENE a JUN 2018

2. Además, de la figura 18 se observa que las variables ENE_18, FEB_18, MAR_18, ABR_18, MAY_18 y JUN_18 de la base de datos Consumo ENE a JUN 2018, tampoco hay atributos perdidos, ni campos vacíos.

3. Observamos, sin embargo, la siguiente situación con la presencia de valores cercanos a 0 en todas estas variables de los consumos mensuales con al menos el 25% de los datos en todos los estados y todos los usuarios en cada mes de consumo. También se nota un cambio alto entre el valor del consumo del cuartil 75% y su máximo en cada uno de los meses.

4. Por otra parte, la primera característica tradicionalmente etiquetada con “Cliente” ahora es etiquetada con “cliente”, así que para un sistema con sensibilidad a las mayúsculas y minúsculas puede generar errores a la hora de transformar los datos para prepararlos para la analítica descriptiva y predictiva.

5. Los niveles 'SUSPENDIDO', 'CORTADO', 'CORTE EN CAMPO' de la variable categórica 'Estado' mantiene los mismos valores porcentuales del semestre anterior.

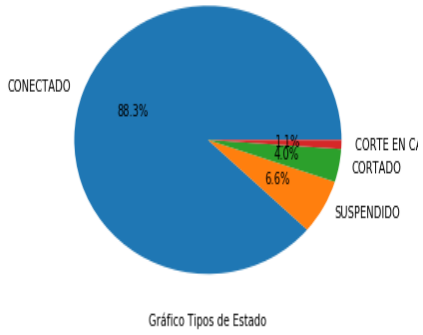


Figura 19. Estado de los usuarios 2018 semestre 1

6. El siguiente diagrama de barras nos muestra una distribución de la identificación de los usuarios. Gráficamente no podemos observar un cambio de la distribución de los usuarios con respecto al semestre anterior.

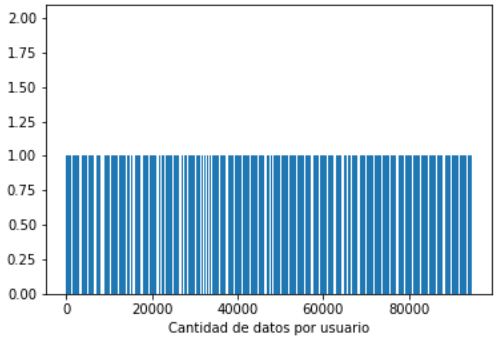


Figura 20. Distribución de la identificación de los usuarios

2.5 Consumo JUL a DIC 2018

1. La base de datos Consumo JUL a DIC 2018 en Excel contiene 8 columnas, donde la primera es el ID del cliente que está en un formato no estándar a recomendaciones en bases de datos, puesto que el ID es dado por un número que aumenta su cantidad de dígitos identificado con la etiqueta con “CLIENTE”. En la columna 8 identificada con “Estado” no tiene niveles faltantes.

	201807	201808	201809	201810	201811	201812
count	89509.000000	89509.000000	89509.000000	89509.000000	89509.000000	89509.000000
mean	147.977180	157.610879	151.901424	161.382533	161.320549	161.425935
std	2799.592137	2977.927163	2872.142255	2875.513892	2997.080859	3118.991185
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	5.000000	5.000000	6.000000	5.000000	7.000000	7.000000
50%	58.000000	60.000000	62.000000	68.000000	69.000000	67.000000
75%	127.000000	133.000000	130.000000	141.000000	140.000000	139.000000
max	663388.000000	659464.000000	658930.000000	626717.000000	653541.000000	686879.000000

Figura 21. Base de datos para Consumo JUL a DIC 2018

2. Dentro de las variables 201807, 201808, 201809, 201810, 201811 y 201812 de la base de datos Consumo JUL a DIC 2018, tampoco hay atributos perdidos, ni campos vacíos, sin embargo, la denominación de estas características se ha denominado diferente al formato anterior. Haciendo una tabla descriptiva de estas variables se mantiene los bajos consumos de energía hasta en un 25% de los usuarios para todos los estados y los meses, también la diferencia entre el cuartil 75% y el valor máximo.

3. Por otra parte, la primera característica tradicionalmente etiquetada con “Cliente” ahora es etiquetada con “CUENTA”, así también la característica etiquetada con “Estado” ahora se identifica con la etiqueta “OBS”. Esto puede generar errores a la hora de transformar los datos para prepararlos para la analítica descriptiva y predictiva.

4. La distribucción histórica de los niveles de la variable permanece practicamente constante, y este semestre se tiene un 11.1% en los niveles diferente al CONECTADO (figura 22).

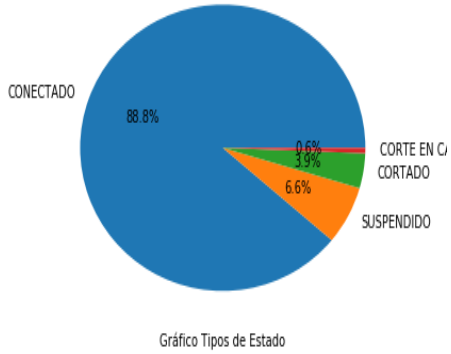


Figura 22. Estado de los usuarios 2018 semestre 2

5. Gráficamente no podemos observar un cambio de la distribución de los usuarios con respecto al semestre anterior y la frecuencia de información es acorde a la situación de medición (figura 23).

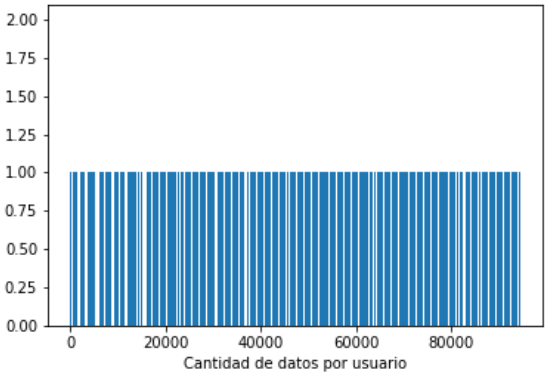


Figura 23. Distribución de la identificación de los usuarios

2.6 Consumo ENE a JUN 2019

1. La base de datos Consumo ENE a JUN 2019 contiene 8 columnas, donde la primera es el ID del cliente que está en un formato no estándar a recomendaciones en bases de datos, puesto que el ID es dado por un número que aumenta su cantidad de dígitos identificado con la etiqueta con “CUENTA”. En la columna 8 identificada con “OBS” no tiene niveles faltantes. Mantuvo el formato anterior del semestre pasado.

	201901	201902	201903	201904	201905	201906
count	89509.000000	89509.000000	89509.000000	89509.000000	89509.000000	89509.000000
mean	156.414495	154.350935	159.923647	151.495335	155.835764	146.285159
std	2633.131817	2734.212312	2857.618562	2732.674453	2951.967864	2845.164441
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	8.000000	8.000000	9.000000	9.000000	10.000000	0.000000
50%	69.000000	66.000000	68.000000	66.000000	66.000000	59.000000
75%	142.000000	135.000000	140.000000	132.000000	134.000000	127.000000
max	559871.000000	557220.000000	626163.000000	600867.000000	663838.000000	640118.000000

Figura 24. Base de datos Consumo ENE a JUN 2019

2. Dentro de las variables 201901, 201902, 201903, 201904, 201905 y J201906 de la base de datos Consumo ENE a JUN 2019, tampoco hay atributos perdidos, ni campos vacíos. Haciendo una tabla descriptiva de estas variables (figura 24) observamos los mismos comportamientos anteriores, y por tanto se necesitará más información de los usuarios para buscar la separación al discriminar usando más características categóricas de los mismos.

3. La distribucción histórica de los niveles de la variable permanece practicamente constante, y este semestre igualmente se tiene un 11.1% en los niveles diferente al CONECTADO.

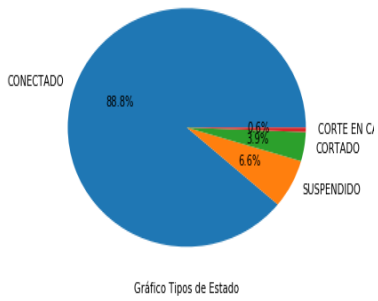


Figura 25. Estado de los usuarios 2019 semestre 1

4. Se mantiene la frecuencia adecuada para cada ID de usuario de la empresa, y una distribución de cuenta, gráficamente, semejante. Salvo los cambios de nombre de algunas variables, diferencias en los valores de consumo, los datos son muy consistentes con el proceso de la adquisición de datos en campo.

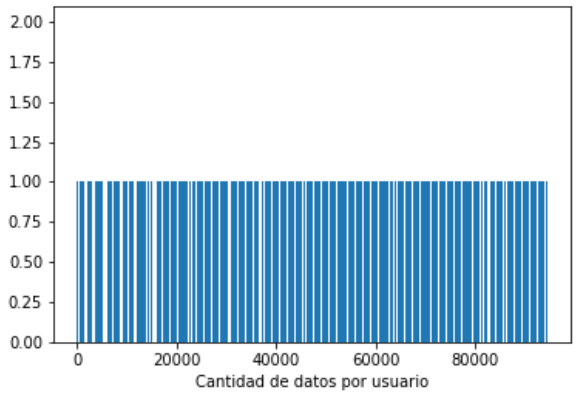


Figura 26. Distribución de la identificación de los usuarios

Por otro lado, al revisar solo en el nivel CONECTADO de la variable Estado (OBS, algunos casos), y cero en todos los valores de los seis meses de consumo en cada base de datos semestral, encontramos 14213 casos de 89657 registros en total, que corresponde al 15.85% aproximadamente del total de los datos de este archivo en el segundo semestre del año 2016.

Realizando el mismo procedimiento con los 6 archivos obtenemos la siguiente tabla.

Tabla 1. *Resumen de conectados con consumos en cero permanente*

	Consumo JUL a DIC 2016	**Consumo ENE a JUN 2017**	**Consumo JUL a DIC 2017**	**Consumo ENE a JUN 2018**	**Consumo JUL a DIC 2018**	**Consumo ENE a JUN 2019**
Usuarios Consumo cero	14213	11756	10315	7759	6057	4293
Total de usuarios	89657	89658	89611	89611	89509	89509
Porcentaje de Usuarios con consumo cero	15,85%	13,11%	11,51%	8,66%	6,77%	4,80%

Aunque hay una disminución de los usuarios del servicio eléctrico de la empresa SETAR, holísticamente se podría asumir una población constante. Bajo este supuesto, la disminución del porcentaje de usuarios que tiene este comportamiento

de consumo cero mientras están conectados al sistema eléctrico de la empresa establece que la empresa ha aumentado sus controles y genera un aumento de los registros con consumo mayor a cero.

3. Estado de las Bases de Datos Espaciales

De acuerdo a las bases de datos espaciales que nos han proporcionado tanto la empresa SETAR como aquella construida por el equipo de trabajo para el distrito 3, después de realizar la salida de campo bajo la orientación de la profesora Roxana, hemos encontrado las siguientes inconsistencias:

1. Los puntos de los medidores que se encuentran en el shapefile del distrito 3 algunos no corresponden a los puntos de los medidores del archivo shapefile de todos los medidores de Tarija (ver figura, puntos morados son distrito tres y los rosa son todo Tarija):

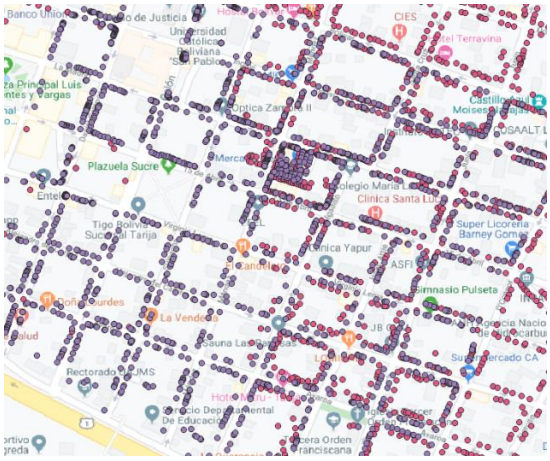


Figura 27. Distribución espacial de los medidores de Tarija

2. Además, de acuerdo a la visualización del gráfico anterior (figura 27), nos damos cuenta que hay medidores que son registrados por el equipo de trabajo (salida de campo) que en archivo vectorial entregado por la empresa no se encuentran.

3. Siguiendo en la misma línea de los hallazgos encontrados y citados previamente, recalcamos sobre las siguientes necesidades:

   ✓ Para cruzar/unir las bases de datos planas y vectoriales, estas deben tener una columna (variable) en común.

Por ejemplo, en la base de datos del 2016 periodo julio – diciembre, tiene una columna denominada Id del cliente (exactamente CLIENTE), pero en el shapefile del distrito 3 no tiene una columna con ese nombre.

- ✓ Los tamaños (número de medidores) de las bases debe ser la misma.

4. En ambos shapefile existen datos perdidos o faltantes, no sabemos si es una información que no observada o errores de digitación.

4. Piloto distrito 3 - Operativos de revisión de medidores

El operativo de revisión de medidores por SETAR y la Universidad Autónoma Juan Misael Saracho desde el 23 de septiembre al 7 de octubre (2019) en el distrito tres de la ciudad de Tarija, realizado en forma conjunta entre los investigadores de la universidad y la empresa Setar, se desarrolló en base a los siguientes objetivos: Localizar las posibles pérdidas no técnicas e identificar las características físicas de los medidores.

Durante diez dias se pudieron revisar 1984 medidores, habiendo iniciado la actividad el dia 9 de septiembre y concluyó el 20 de octubre de 2019.

Tabla 2. *Tipo de medidor*

Tipo	Frecuencia	Porcentaje
MONOFÁSICO	1875	94,5
TRIFÁSICO	109	5,5
Total	1984	100

Fuente: Elaboración propia

De acuerdo con la información en la tabla 2, la mayor parte de los medidores que se revisaron son monofásicos, alcanzando un 94,5% mientras que los trifásicos un 5,5% en el distrito tres de la ciudad de Tarija.

Tabla 3. *Tipo de suministro*

Tipo	Frecuencia	Porcentaje
BAJA TENSIÓN	1964	99,0
MEDIA TENSIÓN	20	1,0
Total	1984	100,0

Fuente: Elaboración propia

En la revisión de medidores la mayor parte tiene suministro de baja tensión (99%), mientras que un 1% suministro de media tensión (Ver tabla 3).

Tabla 4. *Clase de medidores*

Clase	Frecuencia	Porcentaje
ELECTROMECÁNICO	1472	74,2
ELECTRÓNICO CICLO MÉTRICO	184	9,3
ELECTRÓNICO DIGITAL	328	16,5
Total	1984	100

Fuente: Elaboración propia

En la tabla 4 se puede observar que la mayoría de los medidores en el distrito 3 son electromecánicos 74,2%, luego se tienen electrónico digitales 16,5% y electrónico ciclo métrico un 9,3%.

Tabla 5. *Actividad del lugar de revisión*

Clase	Frecuencia	Porcentaje
OFICINAS	170	8,57
INSTITUTOS	7	0,35
ALOJAMIENTO/HOTEL	17	0,86
BANCO	5	0,25
CAJERO AUTOMÁTICO	3	0,15
CANAL DE TELEVISIÓN	1	0,05
CENTRO COMERCIAL	9	0,45
CLÍNICAS/LABORATORIOS	22	1,11
COLEGIO	8	0,40
COMIDA RÁPIDA	2	0,10
CONSULTORIO	10	0,50
FARMACIA	4	0,20
IGLESIA	3	0,15
TIENDA/NEGOCIOS/COMERCIALES	285	14,36
UNIVERSIDAD	13	0,66
VETERINARIA	2	0,10
DOMICILIOS	1423	71,72
Total	1984	100,00

Fuente: Elaboración propia

Por otra parte, en la Tabla 5 se puede observar que la zona objeto de estudio presenta diferentes actividades lo que determina una mayor cantidad de medidores en establecimientos como tiendas, negocios (14,36%) y oficinas (8,57) y el margen restante se reparte en las demás categorías.

Tabla 6. *Observaciones e irregularidades detectadas en medidores*

Observaciones e irregularidades	Frecuencia	Porcentaje
ACOMETIDA EN MAL ESTADO	82	5%
ALTURA DE DIFICIL LECTURACION	13	1%
CAJA EN MAL ESTADO	52	3%
MEDIDOR OBSOLETO	253	14%
CON PUENTE EXTERNO	1	0%
DESCONEXIÓN POR MORA	44	2%
CONEXIÓN DIRECTA/POSIBLE CONEXIÓN DIRECTA	6	0%
DISCO FRENADO	2	0%
HURTO DE ENERGÍA	14	1%
MEDIDOR ADENTRO/NO ACCESIBLE PARA LA LECTURA	827	46%
MEDIDOR QUEMADO	1	0%
MEDIDOR SIN ENERGIA	4	0%
MEDIDOR SIN PRECINTO EN TAPA BORNERA	61	3%
MEDIDOR SIN TAPA BORNERA	429	24%
PRECINTOS VIOLADOS/ROTOS	8	0%
SIN PRECINTO DE BORNERA Y OBSOLETO/SIN TAPA	13	1%
TOTAL	1810	100%

Fuente: Elaboración propia

Las principales observaciones e irregularidades encontradas se presentan en la tabla 6, y son: Los medidores están adentro (46%) y los medidores no tienen tapa bornera (24%)

Tabla 7. *Tabla Posible hurto detectado*

Hurto	Frecuencia	Porcentaje
Posible hurto de energía	17	,9
Sin posible hurto	1967	99,1
Total	1984	100,0

Fuente: Elaboración propia

Tabla 8. *Tipo de medidor donde se detecta el posible hurto*

Tipo	Frecuencia	Porcentaje
MONOFÁSICO	16	94,1
TRIFÁSICO	1	5,9
Total	17	100,0

Fuente: Elaboración propia

Los posibles casos de hurto detectados son 17, los que se van a analizar a continuación a detalle para determinar las características (ver tabla 8).

Las posibles irregularidades detectadas están en tipos de medidores monofásicos (94,1%) y trifásico (5,9%).

Tabla 9. *Tipo de suministro*

Suministro	Frecuencia	Porcentaje
BAJA TENSIÓN	17	100,0
Total	17	100,0

Fuente: Elaboración propia

Los tipos de suministro en las posibles irregularidades son de baja tensión en un 100%, como se puede observar en la tabla 9.

Tabla 10. *Clase de medidor*

Clase	Frecuencia	Porcentaje
ELECTROMECÁNICO	15	88,2
ELECTRÓNICO CICLO MÉTRICO	2	11,8
Total	17	100,0

Fuente: Elaboración propia

Las posibles irregularidades se presentan en clases de medidores electromecánicos y en electrónicos de ciclo métrico (ver tabla 10).

Tabla 11. *Actividad en el lugar de verificación*

Actividad	Frecuencia	Porcentaje
Colegio	1	5,9
Discoteca	1	5,9
Mercado	1	5,9
Domicilios	14	82,4
Total	17	100,0

Fuente: Elaboración propia

Según lo que muestra la tabla 11, las posibles irregularidades se presentan en su mayoría en 14 domicilios, en el mercado Barrientos Ortuño, en una discoteca y en un colegio.

Tabla 12. *Posible irregularidad*

	Frecuencia	Porcentaje
POSIBLE ROBO DE ENERGIA, PRECINTOS VIOLADOS, CONEXIÓN DIRECTA	17	100,0
Total	17	100,0

Fuente: Elaboración propia

La tabla 12 por su parte, muestra que en los 17 casos se encuentran precintos violados, conexiones directas.

Se pudo identificar geográficamente en el Distrito tres las posibles irregularidades y se muestran en la gráfica siguiente:

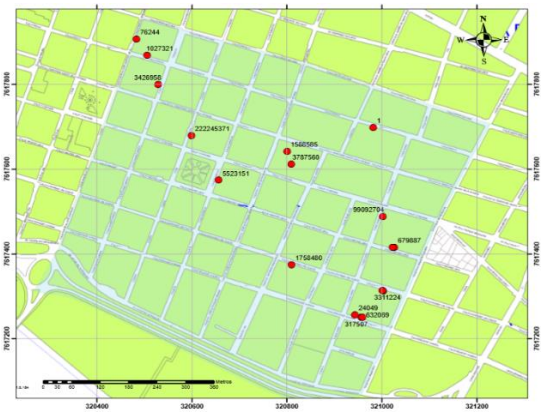


Figura 28. Distribución espacial de los casos irregulares identificados en el Distrito tres

Como parte del análisis, se toma como ejemplo dos casos al azar para analizar el comportamiento en el tiempo y la revisión está relacionada con los reportes hasta diciembre de 2019

Análisis caso: 1687 G1

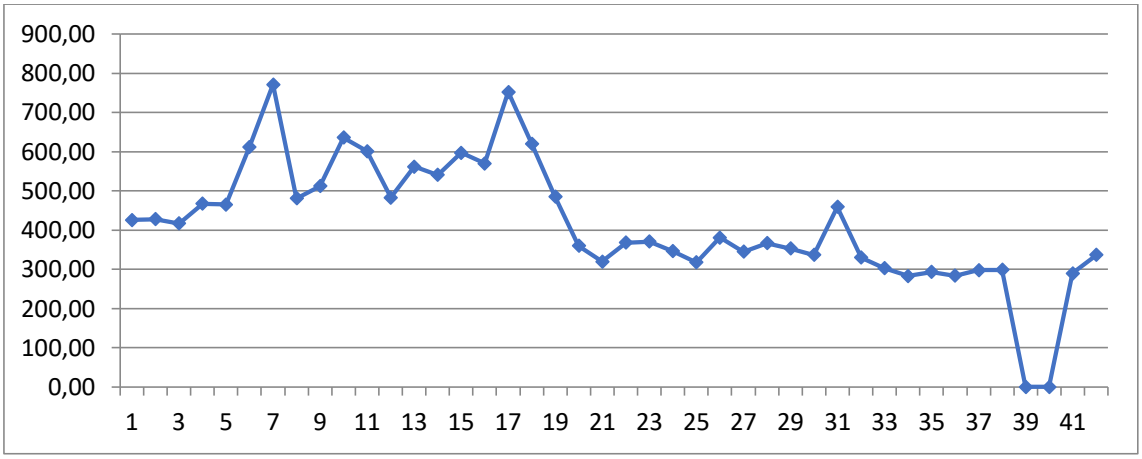


Figura 29. Consumo de energía eléctrica usuario 1687 en 42 meses

En el mes de septiembre y octubre se observa que no hay información de consumo, esta situación requiere de mayor información sobre el caso, por otra parte, al tratarse de un caso con posible irregularidad se puede observar un claro descenso de consumo a partir del mes 21.

Análisis caso: 5789 Domicilio

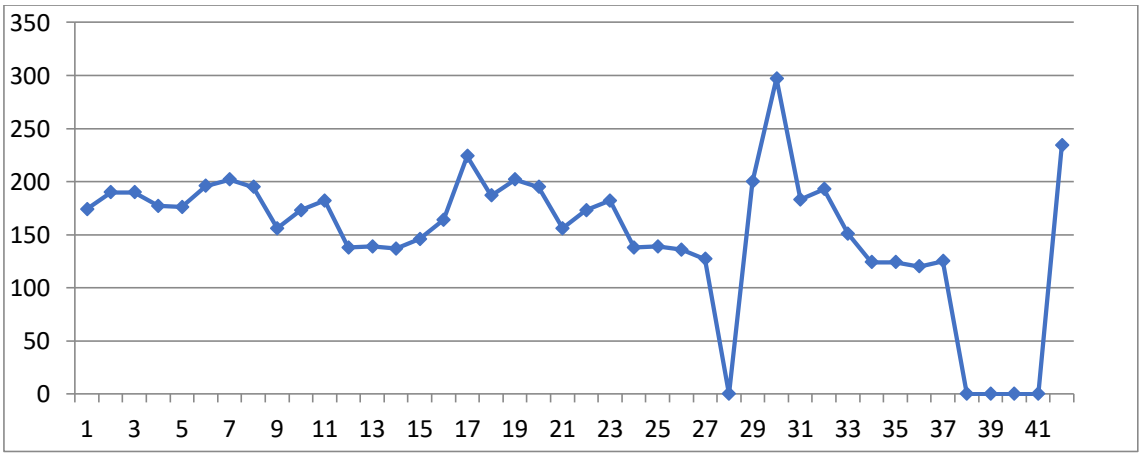


Figura 30. Consumo de energía eléctrica usuario 5789 en 42 meses

Agosto, septiembre, octubre y noviembre no tiene reporte de consumo y se puede observar diciembre presenta un elevado en consumo, en este caso de posible irregularidad se debe indagar las razones del reporte de consumo cero en cuatro meses.

Conclusiones y Recomendaciones

Algunos aspectos importantes a considerar se presentan a continuación:

1. No existe una buena trazabilidad de los datos. Esto implica que la información se tendría que manejar de manera aislada. Por esto, es importante poder garantizar una uniformidad tanto en los nombres de las características en las diferentes datas, como en el código que relaciona los usuarios, de tal manera que se garantice un cruce confiable de la información.

2. La información relacionada con el costo de consumos de los usuarios infractores y no infractores puede colaborar en la detección de las pérdidas no técnicas.

3. Se debe unificar criterios en el registro del nombre de los usuarios.

4. Se recomienda a la empresa, explorar más detalladamente estos usuarios para entender el porqué de este comportamiento de consumo cero cuando están conectados. Esto es una posible causa de pérdidas no técnicas para la empresa.

5. Se requiere mejorar la información de las características socioeconómicas que se tienen registradas de los usuarios de la empresa para iniciar un análisis descriptivo especifico y poder caracterizar los usuarios irregulares, que son los candidatos a convertirse en los usuarios que generan pérdidas no técnicas a la empresa.

6. La reducción en el tiempo de la cantidad de usuarios que reportan un consumo cero semestral, se observa como un reflejo de la intervención que ha realizado la empresa para mejorar los niveles de registros de consumos.

7. El estado de los medidores, así como la accesibilidad a ellos constituyen un factor de riesgo alto para el incremento de pérdidas no técnicas.

8. Dado que el estado actual de las diferentes datas, no tienen relacionada los usuarios irregulares por mes, implica que se desconoce el histórico de estas irregularidades, por lo que se estima que la viabilidad para aplicar las técnicas de minería de datos en la solución del problema planteado es significativamente baja. Por lo tanto se recomienda indagar sobre alternativas adicionales para la solución del problema que se articulen con la data disponible.

9. En la tabla 7, se puede observar la relación entre el número de usuarios infractores y la cantidad total usuarios que tiene la empresa en el distrito tres.

Esta proporción apoya la idea de estudiar un camino alterno para solucionar el problema de las pérdidas no técnicas para la empresa SETAR.

10. Se recomienda que el ID de un usuario que se retire del sistema no sea asignado a otro usuario nuevo, pues esto conduce a ruido en cualquier análisis que se desee realizar sobre la data.

11. Un aspecto importante que agiliza la ubicación de focos de pérdidas no técnicas es el análisis comparativo entre los consumos de energía y lo pagado por ella, de manera tal que se pueda verificar que la energía consumida equivale a la energía pagada. En este sentido se recomienda incluir en la base de datos la información relacionada con la facturación de los usuarios.